

# Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide–protein complexes

Jiro Kondo<sup>1,\*</sup> and Eric Westhof<sup>2</sup>

<sup>1</sup>Department of Materials and Life Sciences, Faculty of Science and Technology, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, 102-8554 Tokyo, Japan and <sup>2</sup>Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire, CNRS, 15 rue René Descartes, 67084 Strasbourg, France

Received April 4, 2011; Revised and Accepted May 16, 2011

## ABSTRACT

**Nucleotide bases are recognized by amino acid residues in a variety of DNA/RNA binding and nucleotide binding proteins. In this study, a total of 446 crystal structures of nucleotide–protein complexes are analyzed manually and pseudo pairs together with single and bifurcated hydrogen bonds observed between bases and amino acids are classified and annotated. Only 5 of the 20 usual amino acid residues, Asn, Gln, Asp, Glu and Arg, are able to orient in a coplanar fashion in order to form pseudo pairs with nucleotide bases through two hydrogen bonds. The peptide backbone can also form pseudo pairs with nucleotide bases and presents a strong bias for binding to the adenine base. The Watson–Crick side of the nucleotide bases is the major interaction edge participating in such pseudo pairs. Pseudo pairs between the Watson–Crick edge of guanine and Asp are frequently observed. The Hoogsteen edge of the purine bases is a good discriminatory element in recognition of nucleotide bases by protein side chains through the pseudo pairing: the Hoogsteen edge of adenine is recognized by various amino acids while the Hoogsteen edge of guanine is only recognized by Arg. The sugar edge is rarely recognized by either the side-chain or peptide backbone of amino acid residues.**

## INTRODUCTION

During almost all biological processes, various proteins recognize nucleic acid molecules. Some of them make tight complexes with DNA [e.g. histone proteins in nucleosome core particles (1)] and RNA [e.g. protein components

in ribosomes (2) and many other ribonucleoproteins (RNPs) (3)], and some bind dynamically and reversibly to nucleic acids [e.g. polymerases (4,5), helicases (6), nucleases (7), transcription factors (8) and aminoacyl-tRNA synthetases (9)]. Proteins such as kinases, G-proteins, motor proteins and chaperones need nucleotides to exhibit their catalytic activities (10,11).

The understanding of the general principles governing nucleic acid recognition by these proteins is therefore necessary to enhance our knowledge of the complex recognition mechanisms underlying all those biological processes. A number of statistical and structural analyses of DNA–protein (12–20) and RNA–protein complexes (20–24) have been performed to determine such recognition principles. Essentially, basic amino acid residues, Arg and Lys, contribute importantly to binding affinity by recognizing negatively charged phosphate backbone of nucleic acids through electrostatic interactions. In addition, various interactions including hydrogen bonds (both direct and water-mediated), C–H...O contacts, van der Waals and cation– $\pi$  interactions increase affinity and specificity in the recognition process. But, each of these interactions separately cannot contribute to sequence-specific recognition because of its geometrical latitude. As Seeman *et al.* (25) proposed in 1976, the use of two hydrogen bonding interactions in the same functional group fixes the position of the two bonds relative to each other and allows one-to-one base-amino acid pairings, such as A-Asn, A-Gln and G-Arg in the major groove (the Hoogsteen edge) of nucleotide bases. Such interactions, called pseudo pairs hereafter, were later found in crystal structures of DNA–protein complexes (15,18). A computational approach taken by Cheng *et al.* (26) found 32 possible pairs (13 pseudo pairs, 19 bifurcated hydrogen bonds where an O, OH or NH<sub>2</sub> group is shared with two acceptor or donor atoms). Of those 32 pairs, 17 (eight pseudo pairs, nine bifurcated

\*To whom correspondence should be addressed. Tel: +81 3 3238 3290; Fax: +81 3 3238 3361; Email: j.kondo@sophia.ac.jp

hydrogen bonds) were indeed observed in DNA/RNA–protein complexes (26).

Here, we have analyzed high-resolution crystal structures of nucleotide–protein complexes in the Protein Data Bank (PDB; <http://www.rcsb.org/pdb>) to update the knowledge of the pseudo pairs composed of two hydrogen bonds, both direct and water-mediated. Other interactions such as electrostatic, C–H...O contacts, van der Waals, cation– $\pi$  and stacking interactions are not considered in this study. For the analysis, we focused on nucleotide–protein complexes over DNA/RNA–protein complexes for two reasons. First, a single nucleotide molecule can bind deeply inside the binding pocket of its target protein and, secondly, the base moiety is free from any other base pairing and can be used for pseudo pairing or hydrogen bonding with amino acid residues at its three interaction edges (Watson–Crick, Hoogsteen, Sugar-edge). In our previous works, RNA base pairs, RNA–ligand base pairs and pseudo pairs were classified by the base edges participating in the interactions (27–30). In this study, a similar classification is applied to base–amino acid pseudo pairs. We also show pseudo pairs from bases to the peptide backbone with two hydrogen-bonding donor (N–H) and acceptor (C=O) groups. These data may be useful not only for our understanding of the molecular recognition in DNA/RNA binding and nucleotide binding proteins but also for peptide and protein engineering.

## MATERIALS AND METHODS

Crystal structures of nucleotide–protein complexes were extracted from the PDB (1 March 2011 release). Since the PDB contains many identical protein structures obtained in different crystallization conditions or those with amino acid mutations, proteins with >30% sequence identity were removed from our data set to minimize redundancy. In this study, a total of 446 structures with resolution better than 2.0 Å, which is 38.6% of structures with resolution better than 3.5 Å, were analyzed manually by using a molecular graphic system *PyMOL* (31) (Supplementary Tables S1–S6). Since crystal structures do not have hydrogen atoms and may contain errors (not only in atomic coordinates and thus on deduced bond distances and angles but also in the choice of amino acid side-chain rotamers), we carefully observed crystal structures one-by-one by eye and picked up hydrogen bonds with a maximum distance of 3.4 Å. The interactions observed between nucleotide bases and amino acids are categorized by hydrogen-bonding patterns, (i) the pseudo pair in which an amino acid side-chain orients in a coplanar fashion and makes at least two hydrogen bonds to a nucleotide base, (ii) the pseudo pair in which peptide backbone atoms of single or multiple residues form at least two parallel hydrogen bonds to a base, (iii) the pseudo pair where an amino acid residue uses both its side- and main-chains to make at least two hydrogen bonds with a parallel arrangement to a base and (iv) the single or bifurcated hydrogen bond. The pseudo pairs corresponding to interactions (i), (ii) and (iii) are shown in Figures 1–4. The single and bifurcated hydrogen bonds corresponding

to interaction (iv) are shown in Supplementary Figures S1 and S2. The numbers of crystal structures at 2-Å resolution containing each of the five bases vary between 63% (A) and 6% (C) (Table 1). There is therefore an overrepresentation of adenine complexes. The number and frequency of each type of pseudo pair and hydrogen bond observed in the present study are summarized in Table 2. The limitation in the number of experimental data sets renders comparisons of hydrogen bonding frequencies between bases difficult. The adenine binding and guanine binding motifs composed of multiple interactions are shown in Figure 5. Molecular drawings were made using *PyMOL* (31).

## RESULTS

### Pseudo pairs and hydrogen bonds

In the present study, all pseudo pairs and hydrogen bonds between nucleotides and amino acids were grouped according to the interaction edges of the bases involved in the pairing (Watson–Crick, Hoogsteen or Sugar-edge) as previously done for other ligands (30). Matrices of pseudo pairs and hydrogen bonds are shown in Figures 1–4 and Supplementary Figures S1 and S2. In Figures 1, 3 and 4, pseudo pairs uniquely classified in this study are boxed by red lines, while those computed by Cheng *et al.* (26) but not observed in this study are boxed by blue lines and those observed both by Cheng *et al.* (26) and us are not boxed. In each panel of the Figures, the nucleotide base, colored in yellow, appears to the left, oriented so that its Watson–Crick edge faces to the right. Amino acid residues are colored in green in these Figures. Each pseudo pair

**Table 1.** The number of crystal structures of nucleotide–protein complexes in PDB

Bases	Nucleotides (Code in PDB)	No. of structures in PDB	
		~ 3.5 Å	~ 2.0 Å (%)
Adenine	Total	739	281 (38.0)
	ATP	238	84 (35.3)
	ADP	343	122 (35.6)
	AMP	158	75 (47.5)
Guanine	Total	174	57 (32.8)
	GTP	80	16 (20.0)
	GDP	87	38 (43.7)
	GMP	7	3 (42.9)
Uracil	Total	96	46 (47.9)
	UTP	16	7 (43.8)
	UDP	65	31 (47.7)
	UMP	15	8 (53.3)
Thymine	Total	82	37 (45.1)
	dTTP (TTP)	45	11 (24.4)
	dTDP (TYD)	16	11 (68.8)
	dTMP (TMP)	21	15 (71.4)
Cytosine	Total	63	25 (39.7)
	CTP	27	8 (29.6)
	CDP	14	7 (50.0)
	CMP (C5P)	22	10 (45.5)
Total		1154	446 (38.6)

Our data set used in this study contains a total of 446 crystal structures with resolution better than 2.0 Å.

**Table 2.** Frequency of each pseudo pair and single hydrogen bond contact observed in our data set of nucleotide–protein complexes

Bases (No. of interactions)	Interaction edges (No. of interactions)	Amino acids	Pseudo pair		Hydrogen bond		
			Number	Frequency (%)	Number	Frequency (%)	
Adenine (266)	Watson–Crick (181)	Asn	10 (6 WM)	3.8 (2.3)	4	1.5	
		Gln	2	0.8	4	1.5	
		Asp	7 (6 WM, 1 Asp/PB)	2.6 (2.3, 0.4)	10	3.8	
		Glu	1	0.4	3	1.1	
		Ser	1 (1 Ser/PB)	0.4 (0.4)	6	2.3	
		Thr	–	–	4	1.5	
		Thy	–	–	4	1.5	
		PB	108 (6 WM)	40.6 (2.3)	17	6.4	
		Hoogsteen (84)	Asn	8	3.0	3	1.1
			Gln	8	3.0	5	1.9
	Asp		1 (1 WM)	0.4 (0.4)	–	–	
	Glu		1 (1 WM)	0.4 (0.4)	19	7.1	
	His		–	–	2	0.8	
	Ser		–	–	1	0.4	
	Thr		–	–	5	1.9	
	Tyr		–	–	2	0.8	
	PB		–	–	29	10.9	
	Ser		–	–	1	0.4	
	Guanine (82)	Sugar-edge (1)	–	–	–	–	
		Watson–Crick (64)	Asn	–	–	1	1.2
Gln			–	–	2	2.4	
Asp			28	34.1	2	2.4	
Glu			2	2.4	7	8.5	
Ser			–	–	14	17.1	
PB			6	7.3	2	2.4	
–			–	–	14	17.1	
Hoogsteen (17)		Asn	–	–	–	–	
		Arg	2	2.4	–	–	
		Thr	–	–	1	1.2	
Uracil (47)		Sugar-edge (1)	–	–	–	–	
		Watson–Crick (38)	PB	–	–	1	1.2
			Asn	2	4.3	–	–
	Gln		1	2.1	–	–	
	Asp		–	–	7	14.9	
	Ser		1 (1 Ser/PB)	2.1 (2.1)	1	2.1	
	Thr		2 (2 Thr/PB)	4.3 (4.3)	–	–	
	Tyr		–	–	1	2.1	
	PB	22	46.8	1	2.1		
	Hoogsteen (4)	Asn	–	–	1	2.1	
		Arg	–	–	2	4.3	
		His	–	–	1	2.1	
		–	–	–	–	–	
	Sugar-edge (5)	Asn	3	6.9	–	–	
		Arg	–	–	1	2.1	
		His	–	–	1	2.1	
		–	–	–	–	–	
–		–	–	–	–		
Thymine (24)	Watson–Crick (17)	Asn	2	8.3	2	8.3	
		Gln	4	16.7	1	4.2	
		Asp	–	–	1	4.2	
		Glu	–	–	1	4.2	
		Lys	–	–	1	4.2	
		PB	4	16.7	1	4.2	
		–	–	–	–	–	
	Hoogsteen (7)	Arg	–	–	5	20.8	
		Trp	–	–	1	4.2	
		PB	–	–	1	4.2	
	Cytosine (24)	Sugar-edge (0)	–	–	–	–	
		Watson–Crick (18)	Asn	1 (1WM)	4.2 (4.2)	–	–
			Arg	3	12.5	–	–
			Ser	–	–	3	12.5
Thr			1 (1 Thr/PB)	4.2 (4.2)	–	–	
PB			6 (1 WM)	25.0 (4.2)	4	16.7	
Hoogsteen (6)		Asp	–	–	1	4.2	
		PB	–	–	5	20.8	
Sugar-edge (0)		–	–	–	–		

The water-mediated pseudo pair is designated as WM. The pseudo pair using both side- and main-chains is designated as Xxx/PB (Xxx = three-letter code of amino acid).

and hydrogen bond from a base to an amino acid residue is named by the interaction edge of the base, where bases and amino acid side-chains are respectively notated by one-letter and three-letter codes and the peptide backbone is abbreviated to PB (e.g. Hoogsteen G-Asp, Watson-Crick A-PB). When an amino acid residue uses both its side- and main-chain for pseudo pairing, the residue is designated as Xxx/PB (e.g. Watson-Crick A-Asp/PB).

#### Amino acid residues participating in pseudo pairs

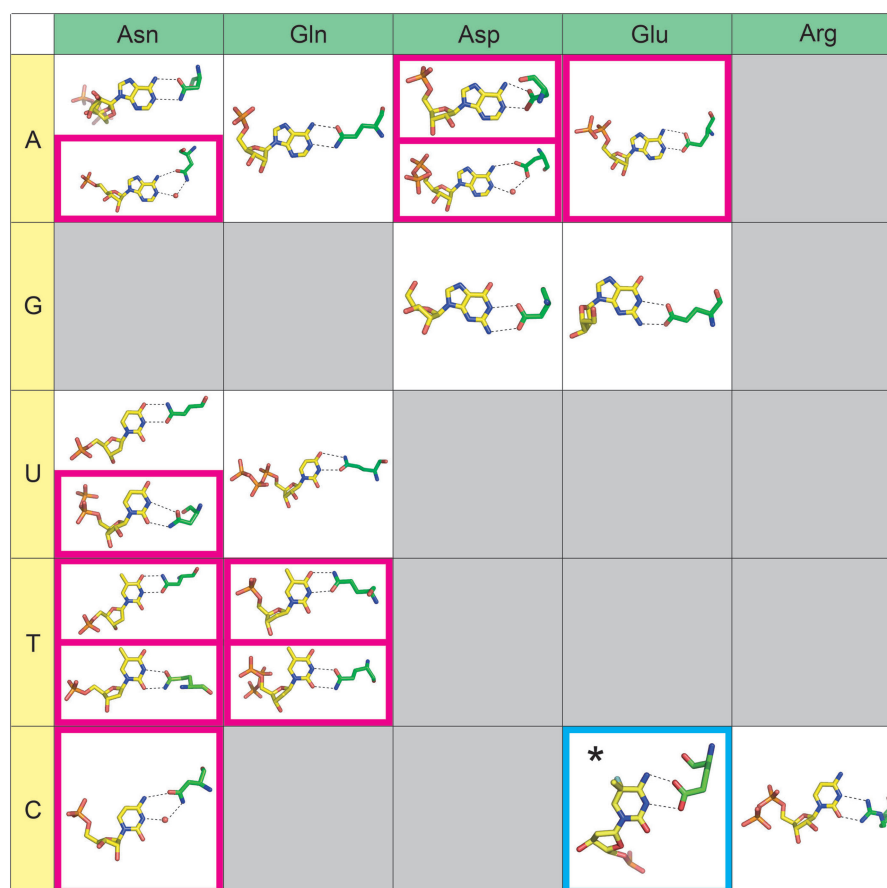
Among the twenty usual amino acids, two polar uncharged ones (Asn, Gln), two acidic ones (Asp, Glu) and a basic one (Arg) possess planar structures with hydrogen-bonding donor and/or acceptor atoms resembling those present in the nucleotide bases. Therefore, they are able to orient in a coplanar fashion and to form pseudo pairs with nucleotide bases through two hydrogen bonds. Other amino acid side-chains can make single and bifurcated hydrogen bonds to nucleotide bases. The peptide backbone also has hydrogen-bonding donor (N-H) and acceptor (C=O) groups and is able to form pseudo pairs and hydrogen bonds. In addition, an acidic (Asp) and two polar uncharged amino acids (Ser, Thr) can make pseudo pairs by using their both side-chain and peptide backbone.

## DISCUSSION

### Watson-Crick pseudo pairs with amino acid side-chains

As expected, the Watson-Crick pseudo pair is the most frequently observed family in nucleotide-protein interactions. Among a total of 17 pseudo pairs, three water-mediated pseudo pairs are observed in this study (Figure 1).

Adenine forms the Watson-Crick pseudo pairs with Asn, Gln, Asp and Glu. In the Watson-Crick pseudo pairs A-Asn and A-Gln,  $N_1(A) \dots H-N(Asn/Gln)$  and  $N_6-H(A) \dots O(Asn/Gln)$  hydrogen bonds are formed. A similar geometry is found in the Watson-Crick A-Asp and A-Glu pseudo pairs, where the  $N_1$  atom of adenine or the  $COO^-$  group of Asp/Glu has to be protonated [the  $pK_a$  values of  $N_1$  in Adenine,  $COO^-$  in Asp and Glu are about 4.0, 3.9 and 4.3, respectively (32,33)]. If Asp and Glu have the resonance-stabilized  $COO^-$  group,  $N_1^+-H(A) \dots O(Asp/Glu)$  and  $N_6-H(A) \dots O(Asp/Glu)$  hydrogen bonds are formed. If these residues have the protonated COOH group,  $N_1(A) \dots H-O(Asp/Glu)$  and  $N_6-H(A) \dots O(Asp/Glu)$  are formed. It is quite probable that the proton is shared between the two groups. For Asn and Asp, water-mediated pseudo pairs with adenine



**Figure 1.** The Watson-Crick pseudo pairs between nucleotide bases and amino acid side-chains. Nucleotides and amino acids are colored in yellow and green, respectively. Water molecules are shown by red spheres. Pseudo pairs classified in this study are boxed by red lines, and those observed by Cheng *et al.* (26) but not in this study are boxed by blue lines. Hydrogen bonds are shown in black dashed lines. In the C-Glu pseudo pairs observed in DNA-protein complexes (PDB-ID = 1dct, 1mht, 4mht), Cytosine modified at position C5 is indicated by an asterisk.



[N<sub>1</sub>(A)...W...H-N(Asn) and N<sub>1</sub>(A)...W...O(Asp) hydrogen bonds] are formed, respectively. These water-mediated pairs (which do not require protonation on either the base or the amino acid) are observed more frequently than the direct pairs; 6 of 10 A-Asn and six of seven A-Asp are the water-mediated pseudo pairs (Table 2). Due to structural similarity, one may expect that Gln and Glu can form identical water-mediated pseudo pairs, but surprisingly such pairs are not observed in this study. Gln and Glu have long side-chains that can reach easily the N<sub>1</sub> atom, which may be one of the reasons why the water-mediated geometries are not observed with these two amino acid residues.

Guanine is recognized by Asp and Glu through the Watson-Crick pseudo pairs with N<sub>1</sub>-H(G)...O(Asp/Glu) and N<sub>2</sub>-H(G)...O(Asp/Glu). The Watson-Crick G-Asp is the major pseudo pair formed by guanine at 34.1% frequency (Table 2). This association was noted by Tregger and Westhof (23). Although Asn and Gln can geometrically make pseudo pairs through N<sub>1</sub>-H(G)...O(Asn/Gln) and O<sub>6</sub>(G)...H-N(Asn/Gln) as proposed by Cheng *et al.* (26), these pairs were not observed either by them or by us.

Both uracil and thymine pair with Asn and Gln in two different geometries at their Watson-Crick edges. In both pseudo-pair geometries, the O atom of CONH<sub>2</sub> makes a hydrogen bond with N<sub>3</sub>-H. On the other hand, the amino group NH<sub>2</sub> can interact either with O<sub>2</sub> or O<sub>4</sub>. However, no U-Gln pseudo pair with O<sub>2</sub>(U)...H-N(Gln) and N<sub>3</sub>-H(U)...O(Gln) hydrogen bonds was found in our data set of nucleotide-protein complexes.

Since cytosine has a nitrogen and an amino group at Positions 3 and 4, respectively, it may be possible to form the Watson-Crick pseudo pairs with Asn and Gln through N<sub>3</sub>(C)...H-N(Asn/Gln) and N<sub>4</sub>-H(C)...O(Asn/Gln). When the N<sub>3</sub> of cytosine [pK<sub>a</sub> is about 4.2 (32)] or COO<sup>-</sup> of Asp/Glu is protonated, the Watson-Crick C-Asp and C-Glu pseudo pairs either with N<sub>3</sub>-H(C)...O(Asp/Glu) or N<sub>3</sub>(C)...H-O(Asp/Glu) and N<sub>4</sub>-H(C)...O(Asp/Glu) may be possible. The C-Glu pseudo pairs (cytosine is modified at position C<sub>5</sub> in PDB-ID = 1dct, 1mht, 4mht) were observed in DNA-protein complexes (26). However, only a water-mediated C-Asn with N<sub>3</sub>(C)...W...H-N(Asn) and N<sub>4</sub>-H(C)...O(Asn) hydrogen bonds is observed in this study. Cytosine is the only base observed forming the Watson-Crick base pair with Arg. The C-Arg pseudo pair has O<sub>2</sub>(C)...H-N(Arg) and N<sub>3</sub>(C)...H-N(Arg) hydrogen bonds. Since Arg has three donor N atoms (two NH<sub>2</sub> and one NH) at its side-chain, four different geometries are possible in principle.

#### Watson-Crick pseudo pairs with the peptide backbone

Since the peptide backbone has both the hydrogen-bonding donor (N-H) and acceptor (C=O) groups, it can form the Watson-Crick pseudo pairs with the nucleotide bases (Figure 2). All pseudo pairs except A-PB and A-Asp/PB shown in Figure 2 are now classified in this study.

In the Watson-Crick A-PB pseudo pairs, N<sub>1</sub> and N<sub>6</sub>-H of adenine make hydrogen bonds with N-H and C=O of

the peptide backbone, respectively. The A-PB pseudo pair is well known as the adenine binding motif identified by Kobayashi and Go (34) and classified by Denessiouk *et al.* (35-37). Many proteins with different folds and functions share the common interaction motif for adenine recognition (Figure 5). The A-PB pseudo pair is thus observed most frequently (40.6% frequency) in ATP, ADP and AMP binding proteins (Table 2). A similar pairing geometry is found in a C-PB pseudo pair where N<sub>3</sub>(C)...H-N(PB) and N<sub>4</sub>-H(C)...O(PB) hydrogen bonds are observed. Cytosine can form other types of pseudo pairs with the peptide backbone, where direct and/or water-mediated hydrogen bonds are observed between O<sub>2</sub>(C) and H-N(PB) and between N<sub>4</sub>-H(C) and O(PB). In the G-PB pseudo pairs, the donor (N<sub>1</sub>-H and N<sub>2</sub>-H) and acceptor (O<sub>6</sub>) groups of guanine are recognized by the acceptor (C=O) and donor (N-H) groups of the peptide backbone, respectively. In both U-PB and T-PB pseudo pairs, the acceptor (O<sub>2</sub> and O<sub>4</sub>) and donor (N<sub>3</sub>-H) groups in bases make hydrogen bonds with the donor (N-H) and acceptor (C=O) groups in the peptide backbone, respectively. The U-PB pseudo pair is the major interaction for uracil recognition with a 46.8% frequency (Table 2).

An acidic (Asp) and two polar uncharged amino acids (Ser, Thr) can make Watson-Crick pseudo pairs by using both the side-chain and peptide backbone atoms. Side-chains of these three amino acid residues are short, which may be the main reason why such pseudo pairings can occur. In the present study, the Watson-Crick A-Asp/PB, A-Ser/PB, U-Asp/PB, U-Ser/PB, U-Thr/PB and C-Thr/PB pseudo pairs are observed. The A-Asp/PB pseudo pairs were previously observed by Denessiouk *et al.* (37). In principle, such pseudo pairings are geometrically possible between any of the five bases and those three amino acid residues.

#### Hoogsteen pseudo pairs

For the Hoogsteen pseudo pair, a total of five geometries are observed with two of them water-mediated pseudo pairs (Figure 3). Three pyrimidine bases, uracil, thymine and cytosine, are not observed to form the Hoogsteen pseudo pair.

Adenine forms the Hoogsteen pseudo pairs with Asn and Gln through N<sub>6</sub>-H(A)...O(Asn/Gln) and N<sub>7</sub>(A)...H-N(Asn/Gln) hydrogen bonds. These two pairs are the most frequent Hoogsteen pseudo pairs taken by adenine (Table 2). In addition, two water-mediated Hoogsteen pseudo pairs, A-Asp and A-Glu, are observed in this study. A water molecule bridges between N<sub>7</sub>(A) and O(Asp/Glu) in these pseudo pairs.

Guanine makes the Hoogsteen pseudo pair only with Arg. Two hydrogen bonds, O<sub>6</sub>(G)...H-N(Arg) and N<sub>7</sub>(G)...H-N(Arg), exist in this pseudo pair. As mentioned above, Arg has three donor N atoms and is possible to form four different pseudo-pair geometries. Although the pseudo pair is frequently observed in DNA-protein complexes (26), only two examples (2.4% frequency) are observed in our nucleotide-protein data set (Table 2).

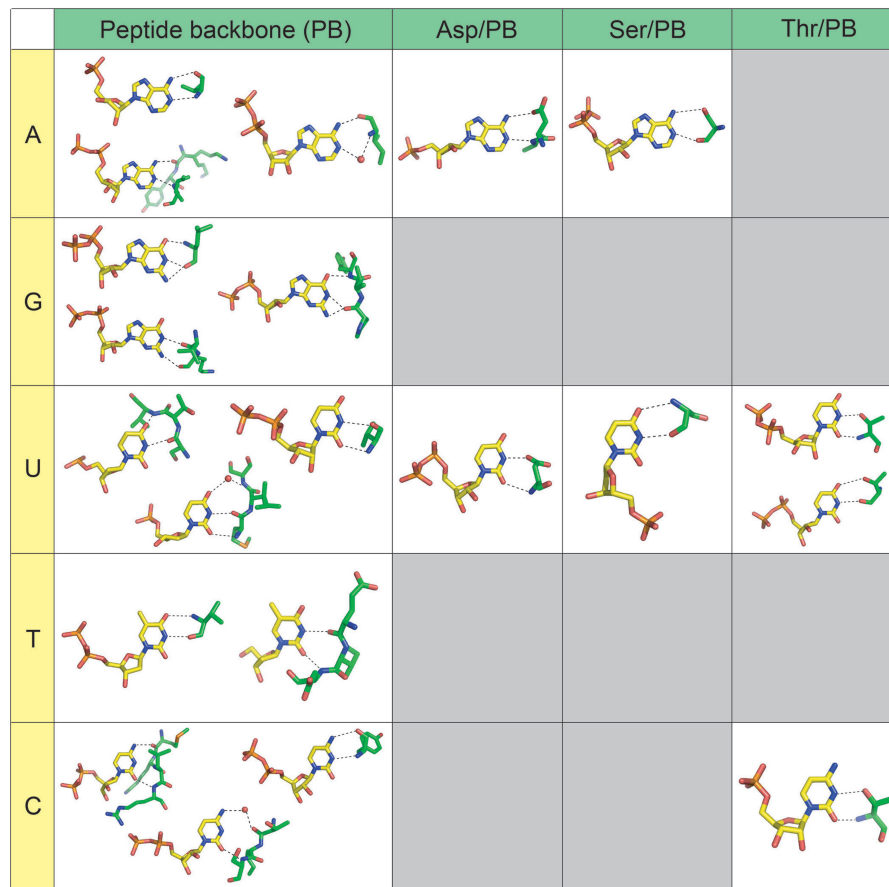


Figure 2. The Watson-Crick pseudo pairs between nucleotide bases and the peptide backbone.

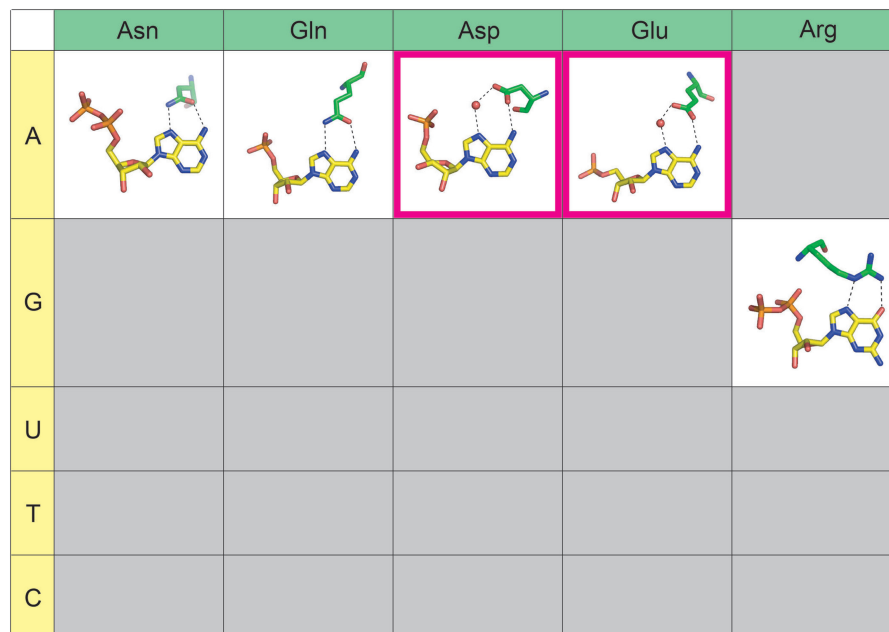


Figure 3. The Hoogsteen pseudo pairs between nucleotide bases and amino acid side-chains. Pseudo pairs classified in this study are boxed with red lines.

The Hoogsteen edge of nucleotide bases is a good discriminatory element for the recognition of nucleotide bases by protein side chains, since the edge of adenine can make pseudo pairs with Asn, Gln, Asp and Glu, while that of guanine pairs only with Arg, and that of pyrimidine bases cannot make pseudo pairs with any of these five amino acid residues.

### Sugar-edge pseudo pairs

The Sugar edge is rarely recognized by amino acid residues through two hydrogen bonds (Figure 4 and Table 2). Obviously, the Watson–Crick and Hoogsteen edges have more discriminatory power and are more efficiently recognized by amino acid residues of proteins.

Among the five nucleotide bases observed to form pseudo pairs, only guanine has two atoms at its Sugar edge that function as hydrogen-bonding donor and acceptor groups. Therefore, two polar uncharged amino acids, Asn and Gln, are able to form the Sugar-edge pseudo pairs through  $N_2-H(G)\dots O(\text{Asn/Gln})$  and  $N_3(G)\dots H-N(\text{Asn/Gln})$  hydrogen bonds as observed by Cheng *et al.* (26). However, these two pseudo pairs are not observed in the present analysis of nucleotide–protein complexes.

Interestingly, a Sugar-edge pseudo pair, in which the  $O_2'$  atom in ribose is used for hydrogen bonding, is observed between U and Asn. This type of pseudo pairing can occur for any RNA bases when the ribose ring has a  $C_2'$ -exo/ $C_3'$ -endo pucker conformation.

### Single and bifurcated hydrogen bonds

The single and bifurcated hydrogen bonds observed in this study are shown and listed in Supplementary Figures S1, S2 and Table 2. The Watson–Crick side of nucleotide base

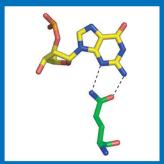
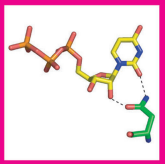
is the major interaction edge participating in single and bifurcated hydrogen bonds as observed for pseudo pairs, since the edge is basically not occupied by any base pairings in nucleotide–protein complexes. On the other hand, the Sugar edge is the minor interaction edge.

Four polar uncharged (Asn, Gln, Ser, Thr), two acidic (Asp, Glu), three basic (Arg, His, Lys) and two hydrophobic (Trp, Tyr) amino acid residues form hydrogen bonds with nucleotide bases. Three hydrogen bonds, the Hoogsteen A–Glu, the Watson–Crick G–Ser and the Hoogsteen G–Asn, are frequently observed in our data set of nucleotide–protein complexes (Table 2).

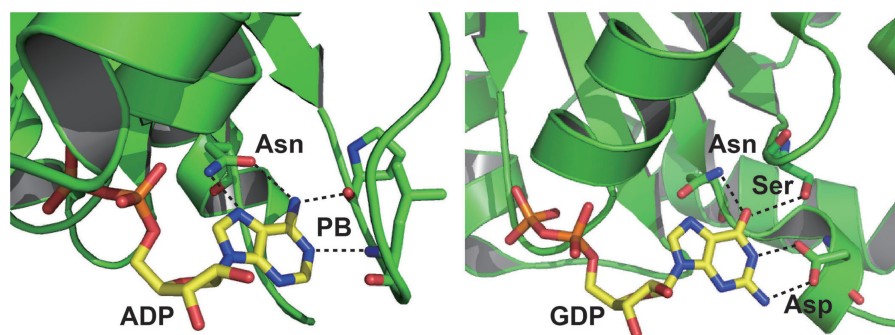
### Multiple interactions make a specific recognition motif

As mentioned above, the adenine binding motif is a common interaction scaffold shared in many ATP, ADP and AMP binding proteins with different folds and functions. In the motif, the Watson–Crick A–PB pseudo pair is the main interaction contributing to the specificity. But adenine is recognized not only through the Watson–Crick edge, but also through the Hoogsteen and Sugar edges (37). An example of the adenine binding motif composed of multiple interactions is shown in Figure 5, where a Hoogsteen A–Asn pseudo pair is observed (38).

A common structural core for GTP/GDP binding exists in G-proteins, and this core includes consensus sequence elements involved in binding of the nucleotide (39). The core of the GTP binding motif consists of five  $\alpha$ -helices and a six-stranded  $\beta$ -sheet, in which five  $\beta$ -strands align in parallel and one in antiparallel. The guanine binding motif is located in a hydrophobic pocket on the surface of the core domain. The guanine base is recognized by multiple interactions containing a Watson–Crick G–Asp pseudo pair, Watson–Crick G–Ser and Hoogsteen G–Asn

	Asn	Gln	Asp	Glu	Arg
A					
G					
U					
T					
C					

**Figure 4.** The Sugar-edge pseudo pairs between nucleotide bases and amino acid side-chains. Pseudo pairs classified in this study are boxed with red lines and those observed by Cheng *et al.* (26) but not in this study are boxed with blue lines.



**Figure 5.** The adenine binding [left; PDB-ID = 1DAD (38)] and guanine binding [right; PDB-ID = 1G7S (40)] motifs composed of multiple interactions. Hydrogen bonds are shown in black dashed lines.

hydrogen bonds, as an example shown in Figure 5 (40). The former pseudo pair apparently contributes to the specificity, which is confirmed by a single point mutation from Asp to Asn that altered the base specificity from GTP to xanthosine triphosphate (41). The latter two hydrogen bonds may have a role for increasing the binding affinity. These three interactions are thus frequently observed in G-proteins at 34.1, 17.1 and 17.1% frequencies, respectively (Table 2).

Such multiple interactions provide both specificity and affinity for ligand recognition. They are observed also in RNA–ligand complexes such as natural riboswitches and synthetic aptamers (30). For example, in the purine riboswitches that regulate translation in response to adenine, guanine, hypoxanthine or 2'-deoxyguanosine, four RNA bases surround all three interaction edges of the purine ligand, encapsulating it completely (42–44). Two of the four bases (residues 74 and 51) are necessary to achieve the high selectivity and two others (residues 22 and 47) contribute to the affinity (45,46). Recently, Dixon *et al.* (47) have successfully developed riboswitches that are selective for synthetic small molecules and no longer respond to the natural intercellular ligands by mutating some of these four bases of the purine riboswitch. It is clear from these observations that specificity and affinity between nucleotides and protein molecules can also be controlled by mutating amino acid residues participating in multiple interactions.

## CONCLUSIONS

A total of 446 crystal structures of nucleotide–protein complexes provide 18 types of direct pseudo pairs and five types of water-mediated pseudo pairs between nucleotide bases and amino acid side-chains. Compared with the previously observed pseudo pairs in DNA/RNA–protein complexes (26), eight direct and five water-mediated new types of pseudo pairs are clustered in our data set of nucleotide–protein complexes. In addition, several pseudo pairs between bases and the peptide backbone are observed in this study.

As expected, pseudo pairs involving the Watson–Crick edge are frequently observed in nucleotide–protein complexes (in which any nucleotide edge is free from

any base pairing). This suggests that the formation of the Watson–Crick pseudo pairs is key for nucleotide selectivity and probably leads to the most stable pseudo pairs. From the frequency of pseudo pairs shown in Table 2, it is clear that the favored pairing partner of adenine and guanine are the peptide backbone and Asp, respectively. The Hoogsteen edge of adenine forms pseudo pairs with Asn, Gln, Asp and Glu. However, the Hoogsteen edge of guanine pairs only with Arg and that of pyrimidine bases was not observed to form any pseudo pair. This suggests that the Hoogsteen edge can discriminate between amino acid side-chains. However, the Sugar edge rarely participates in recognition of amino acid residues. The preference for the Watson–Crick edge compared to the other two edges was already noticed in an analysis of the pseudo-pairs formed between RNA molecules and small molecular ligands (30). However, in the present analysis, the Hoogsteen edge presents a discriminatory power that was not observed for small ligand binding to RNAs. The next step will consist in analyzing complexes formed between proteins and large RNA fragments or molecules.

In this study, matrices of pseudo pairs and hydrogen bonds between nucleotide bases and amino acids are updated. Although our data set is limited to high-resolution crystal structures of nucleotide–protein complexes, some preferences between pseudo pairs and hydrogen bonds in such complexes are revealed by the frequency data. Conclusions based on frequencies should be mitigated by the limited size of the data set (in part dictated by biology itself since, for example, adenine nucleotides are ubiquitously used in biochemical processes). It is still striking to observe the infrequent use of the Sugar edge, which is so typical of RNA (use of the hydroxyl O<sub>2</sub>') and so prevalent in RNA–RNA interactions (A-minor contacts) (29). Another remark concerns the use of a water molecule to mediate a contact between two acceptor atoms instead of base or side chain protonation (or proton sharing between them) (Figure 1). Interestingly, only water-mediated contacts are observed between the acidic groups of Asp and Glu and the Hoogsteen edge of adenine. The  $pK_a$  of N<sub>7</sub>(A) is below 2 and, thus, because of the large difference with that of the acidic group, would not lead to a proton sharing as is the case



with N<sub>1</sub>(A) or N<sub>3</sub>(C). In any case, the data may be useful not only for understanding of the nucleosides, nucleotides and nucleic acids recognitions by proteins, but also for structure-based peptide/protein engineering and drug design.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## FUNDING

Funding for open access charge: Sophia University, Tokyo, Japan.

*Conflict of interest statement.* None declared.

## REFERENCES

- Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.
- Noller, H.F. (2005) RNA structure: reading the ribosome. *Science*, **309**, 1508–1514.
- Nagai, K. (1996) RNA-protein complexes. *Curr. Opin. Struct. Biol.*, **6**, 53–61.
- Brautigam, C.A. and Steitz, T.A. (1998) Structural and functional insights provided by crystal structures of DNA polymerases and their substrate complexes. *Curr. Opin. Struct. Biol.*, **8**, 54–63.
- Werner, M., Thuriaux, P. and Soutourina, J. (2009) Structure-function analysis of RNA polymerases I and III. *Curr. Opin. Struct. Biol.*, **19**, 740–745.
- Caruthers, J.M. and McKay, D.B. (2002) Helicase structure and mechanism. *Curr. Opin. Struct. Biol.*, **12**, 123–133.
- Dupreux, C.M. (2008) Roles of metal ions in nucleases. *Curr. Opin. Chem. Biol.*, **12**, 250–255.
- Latchman, D.S. (1997) Transcription factors: an overview. *Int. J. Biochem. Cell. Biol.*, **29**, 1305–1312.
- Beuning, P.J. and Musier-Forsyth, K. (1999) Transfer RNA recognition by aminoacyl-tRNA synthetases. *Biopolymers*, **52**, 1–28.
- Vetter, I.R. and Wittinghofer, A. (1999) Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer. *Q. Rev. Biophys.*, **32**, 1–56.
- Leipe, D.D., Wolf, Y.I., Koonin, E.V. and Aravind, L. (2002) Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.*, **317**, 41–72.
- Pabo, C.O. and Sauer, R.T. (1984) Protein-DNA recognition. *Annu. Rev. Biochem.*, **53**, 293–321.
- Matthews, B.W. (1988) Protein-DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
- Brennan, R.G. and Matthews, B.W. (1989) Structural basis of DNA-protein recognition. *Trends Biochem. Sci.*, **14**, 286–290.
- Pabo, C.O. and Sauer, R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.
- Suzuki, M. (1994) A framework for the DNA-protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*, **2**, 317–326.
- Mandel-Gutfreund, Y., Schueler, O. and Margalit, H. (1995) Comprehensive analysis of hydrogen bonds in regulatory protein-DNA-complexes: in search of common principles. *J. Mol. Biol.*, **253**, 370–382.
- Luscombe, N.M., Laskowski, R.A. and Thornton, J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
- Coulocheri, S.A., Pigis, D.G., Papavassiliou, K.A. and Papavassiliou, A.G. (2007) Hydrogen bonds in protein-DNA complexes: where geometry meets plasticity. *Biochimie*, **89**, 1291–1303.
- Nadassy, K., Wodak, S.J. and Janin, J. (1999) Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
- Jones, S., Daley, D.T., Luscombe, N.M., Berman, H.M. and Thornton, J.M. (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.
- Allers, J. and Shamoo, Y. (2001) Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.*, **311**, 75–86.
- Treger, M. and Westhof, E. (2001) Statistical analysis of atomic contacts at RNA-protein interfaces. *J. Mol. Recognit.*, **14**, 199–214.
- Morozova, N., Allers, J., Myers, J. and Shamoo, Y. (2006) Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*, **22**, 2746–2752.
- Seeman, N.C., Rosenberg, J.M. and Rich, A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
- Cheng, A.C., Chen, W.W., Fuhrmann, C.N. and Frankel, A.D. (2003) Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J. Mol. Biol.*, **327**, 781–796.
- Leontis, N.B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–612.
- Leontis, N.B., Stombaugh, J. and Westhof, E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
- Stombaugh, J., Zirbel, C.L., Westhof, E. and Leontis, N.B. (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.*, **37**, 2294–2312.
- Kondo, J. and Westhof, E. (2010) Base pairs and pseudo pairs observed in RNA-ligand complexes. *J. Mol. Recognit.*, **23**, 241–252.
- DeLano, W.L. (2008) *The PyMOL Molecular Graphics System*. DeLano Scientific LLC, Palo Alto, CA, USA.
- Saenger, W. (1984) *Principles of Nucleic Acid Structure*. Springer-Verlag, New York.
- Schult, G.E. and Schirmer, R.H. (1979) *Principles of Protein Structure*. Springer-Verlag, New York.
- Kobayashi, N. and Go, N. (1997) A method to search for similar protein local structures at ligand binding sites and its application to adenine recognition. *Eur. Biophys. J.*, **26**, 135–144.
- Denessiouk, K.A. and Johnson, M.S. (2000) When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins*, **38**, 310–326.
- Denessiouk, K.A., Rantanen, V.V. and Johnson, M.S. (2001) Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. *Proteins*, **44**, 282–291.
- Denessiouk, K.A. and Johnson, M.S. (2003) “Acceptor-donor-acceptor” motifs recognize the Watson-Crick, Hoogsteen and Sugar “donor-acceptor-donor” edges of adenine and adenosine-containing ligands. *J. Mol. Biol.*, **333**, 1025–1043.
- Huang, W., Jia, J., Gibson, K.J., Taylor, W.S., Rendina, A.R., Schneider, G. and Lindqvist, Y. (1995) Mechanism of an ATP-dependent carboxylase, dethiobiotin synthetase, based on crystallographic studies of complexes with substrates and a reaction intermediate. *Biochemistry*, **34**, 10985–10995.
- Kjeldgaard, M., Nyborg, J. and Clark, B.F. (1996) The GTP binding motif: variations on a theme. *FASEB J.*, **10**, 1347–1368.
- Roll-Mecak, A., Cao, C., Dever, T.E. and Burley, S.K. (2000) X-Ray structures of the universal translation initiation factor eIF5B: conformational changes on GDP and GTP binding. *Cell*, **103**, 781–792.
- Weijland, A. and Parmeggiani, A. (1993) Toward a model for the interaction between elongation factor Tu and the ribosome. *Science*, **259**, 1311–1314.
- Serganov, A., Yuan, Y.R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A.T., Hobartner, C., Micura, R., Breaker, R.R. and Patel, D.J. (2004) Structural basis for discriminative regulation

- of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.*, **11**, 1729–1741.
43. Batey, R.T., Gilbert, S.D. and Montange, R.K. (2004) Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature*, **432**, 411–415.
44. Edwards, A.L. and Batey, R.T. (2009) A structural basis for the recognition of 2'-deoxyguanosine by the purine riboswitch. *J. Mol. Biol.*, **385**, 938–948.
45. Gilbert, S.D., Stoddard, C.D., Wise, S.J. and Batey, R.T. (2006) Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. *J. Mol. Biol.*, **359**, 754–768.
46. Gilbert, S.D., Love, C.E., Edwards, A.L. and Batey, R.T. (2007) Mutational analysis of the purine riboswitch aptamer domain. *Biochemistry*, **46**, 13297–13309.
47. Dixon, N., Duncan, J.N., Geerlings, T., Dunstan, M.S., McCarthy, J.E., Leys, D. and Micklefield, J. (2010) Reengineering orthogonally selective riboswitches. *Proc. Natl Acad. Sci. USA*, **107**, 2830–2835.